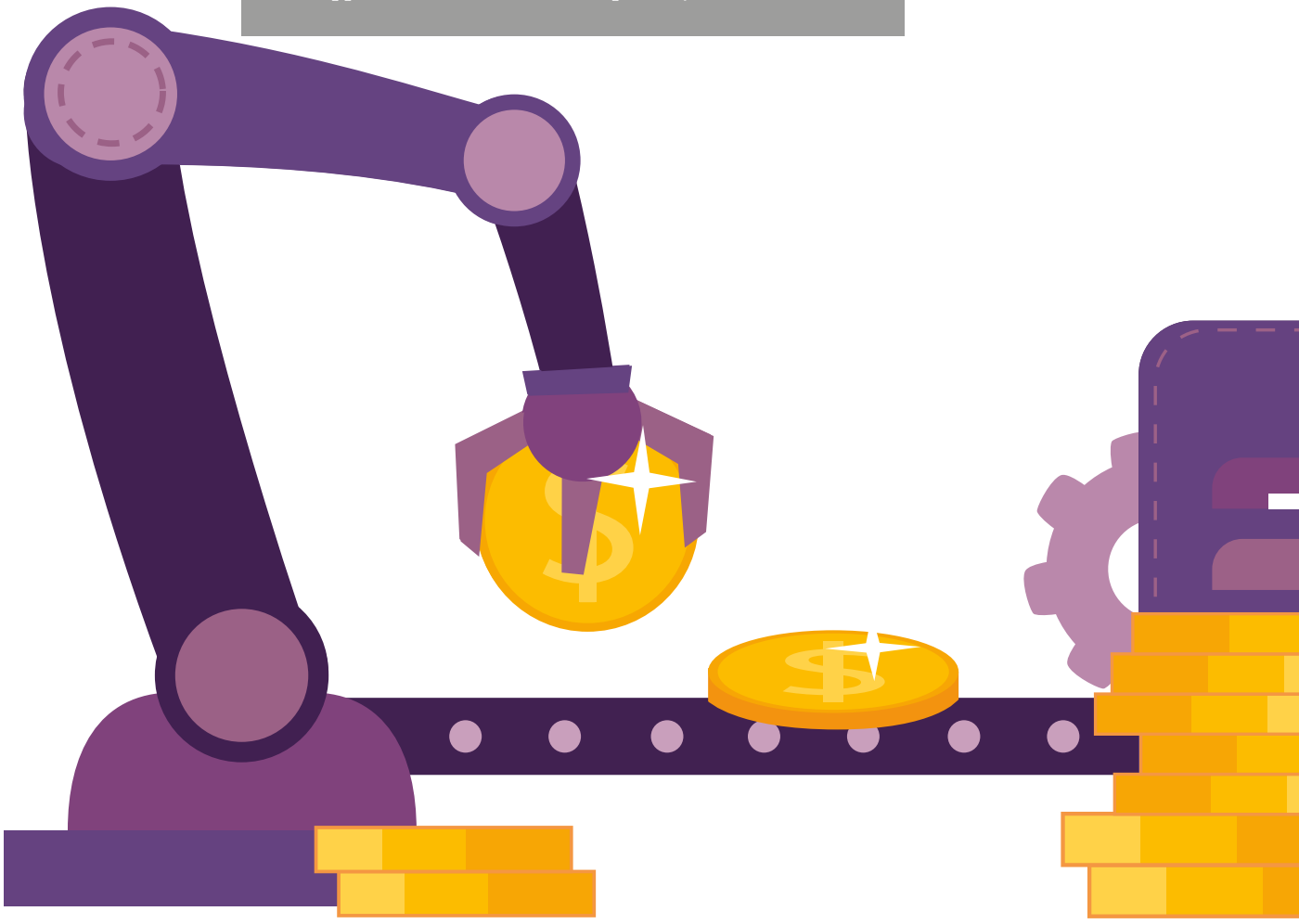
STATISTICS AND AI-ENHANCED AUTOMATION IN BANKING TRANSACTION TESTING

Contents



BY JEFF HULETT, RONALD CATHCART, CONWAY DODGE, AND KRIS MCINTIRE

Background

The use of transaction testing, also known as substantive testing, in banking is a longstanding practice. It is generally a control to ensure transactions are performed correctly. We've seen that, typically, statistics are used early in the transactional testing process when it is determined a sampling approach is appropriate to draw an inference upon the population. Other types of sampling, such as judgmental sampling, are used when an inference is not required for the population. Increasingly, the actual testing is performed using automated tools. Depending on the kind of automation, statistics may also be applied as part of the automation. This is especially true when it is necessary to structure data found on unstructured media.

Transaction testing is generally related to two control types.[1] A preventative control is performed while the transaction is still in process to inform the process operator as to issues that need to be corrected prior to transaction finalization. Alternatively, a detective control is performed after the fact to inform process operators as to risks and process changes that may be necessary as well as signal the potential need for transaction remediation. Statistics and automation are used in both transactional testing types. *(Please note: There are nuances as to how statistics or automation are applied to the different transaction testing types, but we don't detail them in this article.)*

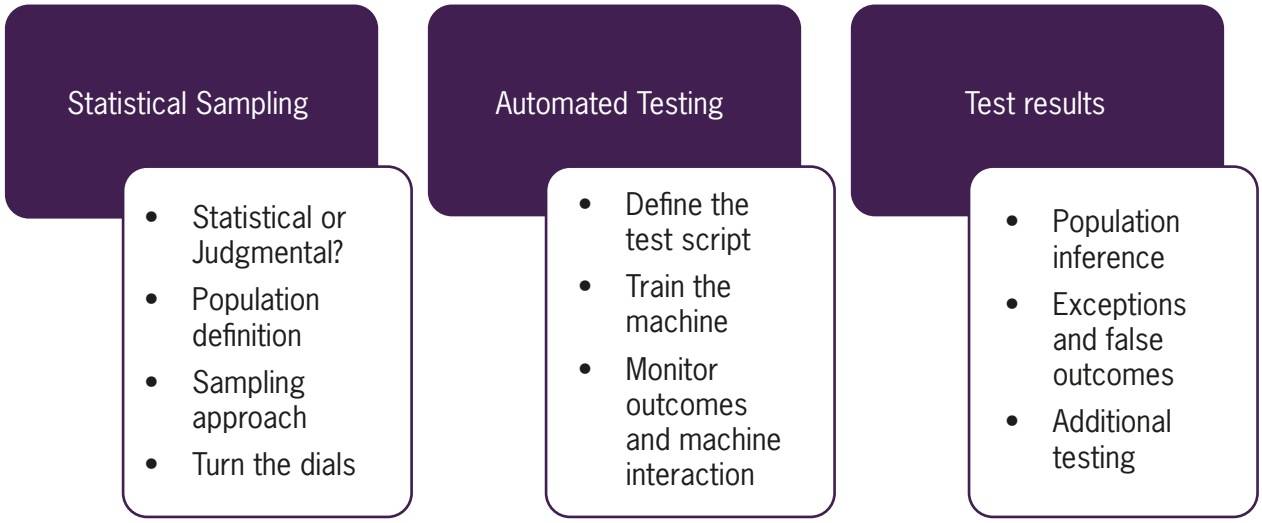
The schematic on the following page shows the high-level testing flow for a statistically enabled testing program.

Statistical Sampling:

Statistics have been traditionally used to determine the sample size of a population necessary to make statistically significant inference upon the population. The difference between statistical and judgmental sampling is that a properly constructed statistical sample will render an inference on the population. On the other hand, judgmental testing is NOT able to render such an inference. Statistical testing has a cost, and many institutions are looking to minimize costs for such testing (for example, a desired inference could be about compliance exceptions found in last year's loan originations population, as inferred from a sample of those originations). To balance this, there are several statistical methods available and adjustment dials to trade off sample size and the precision required for the inference. For example, binary testing (such as pass/fail or exception/no exception) often applies a specific testing approach using a binomial probability mass function.[2] Most important is for the risk reviewer to understand the dials and enough about the underlying math to make an informed testing design decision. These decisions are often made with the assistance of statistical sampling specialists. Section 2 contains more specifics on the statistical sampling methods and approach.[3]

Automation:

Automation technology is a more recent application for transactional testing. Simpler edit checks based on structured data have been around for decades. However, more recent technological innovations provide the use of automation to transform unstructured data to structured data for obligation testing. For example, using artificial intelligence (AI), natural language processing (NLP), and optical character recognition (OCR) enabled technologies, risk reviewers can structure the unstructured data found in documents or in voice recordings. The structured data set can then be tested against compliance rules or other obligations to ensure the transaction is free of compliance exceptions. The technology can perform high volumes of tests, potentially reducing costs associated with people testers. Also,



since high volume and repetitive testing tends to be very boring, the new technology liberates people testers to perform higher cognitive value testing (such as researching exceptions).

Automated testing may have a statistical element to it.[5] For example, the automation engine assigns a probability that the structuring process is correctly completed for each media type (e.g., document image or recording). This is necessary because of the vagaries associated with unstructured media itself (e.g., a lower-resolution image may have a lower-accuracy probability than a higher-resolution image). Also, AI relies on machine learning to interpret the unstructured media. Further, the unstructured media provided for testing may be newer, and thus, the AI less trained. As a result, the unstructured media may be rated as a lower-accuracy probability. Alternatively, the unstructured media may be regularly encountered, resulting in more highly trained AI. In this case, the unstructured media may be rated as a higher-accuracy probability.[6] In general, automation is more effective with higher-volume and relatively homogenous banking product types. *(For a banking product type automation adaptation description, see Appendix 1.)*

Sampling Methodology and Usage

There are two types of statistical sampling methodologies relevant for monitoring and testing activities: sampling for quality testing and sampling for confidence interval estimation.[7]

1. **Quality testing sampling** can be used to assess whether a financial institution's controls and processes are operating with sufficient effectiveness to provide reasonable assurance of quality.
2. **Confidence interval sampling estimation** is suitable when the goal is to determine the sample size that provides an estimate of a population characteristic (e.g., an exception rate) with a predetermined level of precision.

Generally, the objective of compliance testing is to detect problems in a control or process, rather than to provide a precise estimate of the exception rate. Thus, quality testing sampling is used more than confidence interval estimation sampling in compliance testing. In this section, we will reference OCC sampling methodology as related to quality testing.

Defining the Population:

The viability of statistical sampling relies on a well-defined population. If the population is not appropriately specified, the resulting inference based on a population's sample could be flawed (e.g., the inference could lack accuracy). Key accuracy considerations when defining populations include:[8]

- Scope and objectives of the risk reviewer activity.
- Characteristics of the population, and whether the population is homogenous with respect to risk factors.
- Relevant time period. Sample results only apply to the time period that defines the population, when applicable. The external environment should be consistent across the time period.
- The type(s) of exceptions for which the risk reviewers are testing.
- The independence of the population participants.

The Dials - Tolerance and Confidence:

Definition:

- **Tolerance:** The tolerance rate is the limiting rate of exceptions that risk reviewers target not to exceed in the defined population.
- **Confidence:** The confidence level is the level of statistical assurance that conclusions about the defined population based on the sample are accurate.

Determination:

We find that, in general, tolerance and confidence are related to the bank in terms of past risk issues and the strength

On Previous Page: Viktoria Kurpas | Shutterstock.com

			CONFIDENCE		
			Lower Risk ⇨		Higher Risk
			90%	95%	99%
TOLERANCE	Lower Risk	10%	22	29	44
		7%	32	42	64
	⇩	5%	45	59	90
		3%	76	99	152
	Higher Risk	1%	230	299	459

of risk management capabilities. That is, based on past risk issues, the risk reviewer may want a more precise inference. However, there are subtle differences between tolerance and confidence:

- **Tolerance Rate:** This usually contains a historical risk view, considering past issues, the bank's condition, and legal tolerances. In general, higher risk relates to requiring lower tolerance.
- **Confidence Level:** This usually contains a forward view, considering the quality of risk management and ability to handle issues going forward. In general, higher risk relates to requiring a higher confidence level.

Sample size requirements:[9]

This table above describes sample size requirements for a defined population, based on confidence level and tolerance rate requirements. This methodology description addresses the steps that risk reviewers should follow when performing statistical sampling to estimate the population exception rate of a binary attribute (e.g., an outcome of yes/no, true/false, violation/no violation, or exception/no exception). The sample volumes are derived from the binomial probability mass function. Some sampling methodologies also use the population size to determine the necessary sample size. This binomial model does not require population size.[10]

Interpreting the results:

From a statistical standpoint, the results are interpreted with the following built-in assumptions: (1) The target exception rate is 0% and (2) the sampling results are compared to an upper limit based on the assigned tolerance rate and confidence level. *(See Appendix D in the Office of the Comptroller of the Currency Sampling Methodology[11] for the comparison tables.)* An example we have developed: A risk reviewer wants to assess exceptions to certain compliance rules. Based on the defined population, the reviewer randomly chose 29 loans based on 10% tolerance and 95% confidence. Two transactions with compliance exceptions are found. Using Appendix D, the risk reviewer can make the following inference statement: "With 95% confidence, up to 20.16% of the defined population could have compliance exceptions."

Transactional Testing Automation

As a point of emphasis, we've found that a critical step to transactional testing automation success is a rigorous test script definition process. Clarity of risk obligation, regulation, test objective, test steps, and structured and unstructured data sources should be well defined. This is truly a case for the aphorism: "A stitch in time saves nine."

Manual test outcomes	Statistical test outcomes
Pass	True Pass
Fail	True Fail
Exception	Machine Exception
	False Pass
	False Fail

With the overlay of a statistical outcome regarding the testing (as compared to manual testing), additional diligence should be performed to validate the outcome.

False Positives and False Negatives:

False Positives: A false positive error, or false positive, is a result that indicates a given condition exists when it does not. For example, a cancer test that indicates a person has cancer when they do not. This is a Type I error where the test is checking a single condition and wrongly gives an affirmative (positive) decision. However, it is important to distinguish between the Type I error rate and the probability of a positive result being false. The latter is known as the false positive risk.

False Negatives: A false negative error, or false negative, is a test result that wrongly indicates that a condition does not hold.

		Reject null hypothesis	
		No	Yes
Null hypothesis	True	True negative	False positive Type I: α
	False	False negative Type II:β	True positive

For example, when a cancer test indicates a person does not have cancer, but they do. The condition "the person has cancer" holds but the test (the cancer test) fails to realize this condition, and wrongly decides that the person does not have cancer. A false negative error is a Type II error occurring in a test where a single condition is checked for, and the result of the test is erroneous.[12]

Implications:

Depending on the test context the error type has significantly different implications. The cancer example is closest to

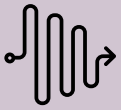
banking transactional testing. A false positive can be annoying or provide the patient/client unnecessary apprehension. A false negative can be deadly; that is, the cancer remains undetected. In the case of bank risk testing, a false positive can create a customer service problem or a false risk signal. A false negative can enable the very risk it is trying to detect. That is, not identifying credit, compliance, or fraud risk when it exists. We've seen that false negatives can be the basis for regulatory enforcement action.[13]

A potential organizational pitfall we've seen is the reaction of reviewed organizations. If a risk organization is testing an operational process, the operations leadership may push back on the results, adding time to resolve the problem. Yet resolving a false positive improves the overall inferred conclusion. We've found that operations leadership is traditionally less active in challenging potential false negatives even though this is where the risk can be more significant. False negative validation, though harder and less organizationally sensitive, is critical.

As an aside, the statistical names false positive or false negative refer to what they are trying to predict. They are not related to the normative judgment of that which is being predicted. For example, in risk testing, a false fail from a risk test is often a false positive and a false pass is often a false negative.

"We've seen that false negatives can be the basis for regulatory enforcement action."

A Compliance Testing Example[14]



Situation: *Customer Communication* – It is necessary to confirm that customers who receive bank communication are receiving what they should under the bank's obligations. There are many federal regulations, like Reg B, Reg Z, and Reg X, state regulations, and investor requirements that create related bank regulatory communication obligations. As such, continuous risk testing needs to be performed.

Complication: Managing customer media channels is operationally complex. There are potentially hundreds of obligations, across multiple products and customer channels, that need to be managed. The obligations change regularly, and the communications are regularly added or changed. Also, third-party vendors are often involved in the customer communication process.

Resolution: Perform automated testing comparing the actual media communication to the obligations. The automation reads the communication, converts it from unstructured to structured data, then uses algorithms to compare the structured data to the obligation rules.

Second Order Complication and Resolution:
A potential automation *false positive*: The automation identifies a document it cannot read or a rule it cannot interpret. Or, the automation identifies a document as a compliance issue, and it is not.

> *Outcome:* There is no risk and the issues are resolved within the bank. Both machine learning and machine confidence are adjusted to manage the false positive.

A potential automation *false negative:* The automation misinterprets data and is not identified by the automation. Or, a rule is not interpreted correctly and does not identify a document as a compliance issue, and it is.

> *Outcome:* The risk is not identified and persists. Human quality control is used as a safety net; this includes selected higher-risk transactions that may require more judgment or have higher severity.

Automated and Statistical Testing Categories and Implications for the Risk Reviewer:

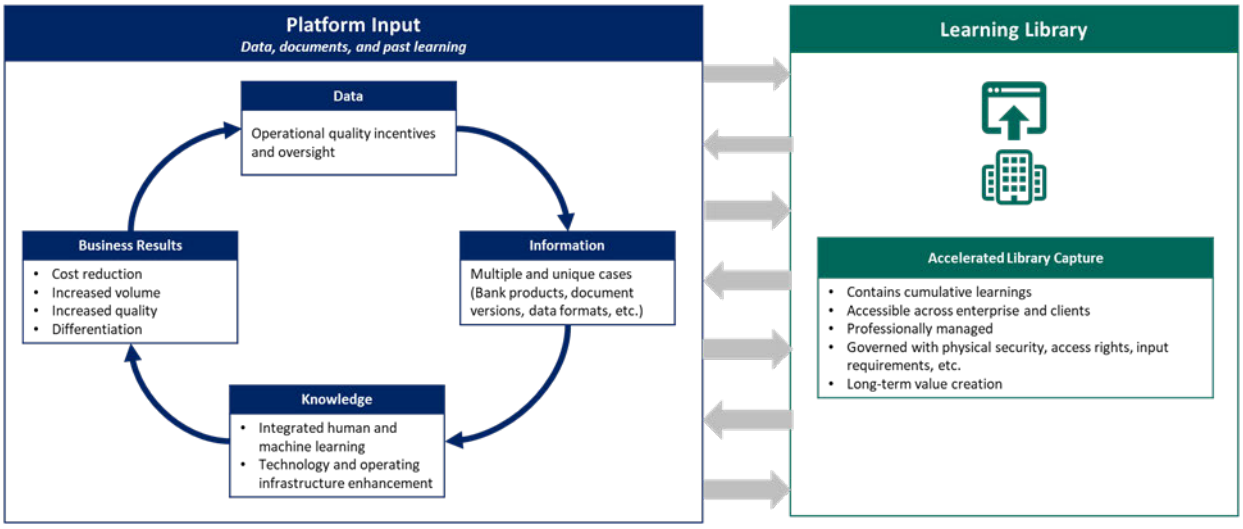
The image to the right shows the five categories for statistically based automated testing and the implications.

- **Machine Awareness:** This relates to whether the machine's coding and learning capability is aware as to whether a transaction is a pass or fail. In addition, an aware machine knows the errors it has been programmed with or has learned to interpret. Errors could include difficulty reading text, ambiguity about a rule, or a media input (document image or voice recording) it has not been trained to understand. In the cases of false passes or false fails, by definition, the machine is not aware. For example, this could occur because the statistical setting was set to a lower accuracy probability on reading an image, and it misinterpreted the image.
- **Human Correction:** This refers to the human activity needed to correct the machine-identified exception. It could be as simple as reading the document and inputting the data the machine does not understand. It could be evaluating the transaction regarding a risk rule the machine is not able to interpret. Our experience is most issues are data related. Generally, if the data is structured properly, the machine should be able to evaluate it via the decision rules.
- **Human Testing:** This is related to uncovering false fails and false passes. Since it is the false pass that is the most dangerous to a bank's risk position, we believe close attention should be given to the transaction the machine believes is a true pass. This can be done by additional testing, usually on a judgmental basis. Similarly, additional testing should be completed on the transaction that the machine believes is a true fail.

Please note: In many statistically based automation tests, there is a statistical tradeoff between false passes and false fails. That is, reducing one category means increasing another

	Category	Machine awareness?	Human correction?	Human testing?	Human operating approach
Five possible statistical test outcomes	True pass	Yes	No	No	Update learning from false resolution
	True fail	Yes	No	No	Update learning from false resolution
	Machine exception	Yes	Yes	No	Patch data and resolve exceptions
	False pass	No	No	Yes	Test passes for accuracy, usually using risk-based sampling and testing
	False fail	No	No	Yes	Test fails for accuracy, usually all or a large portion

The accumulated platform learning creates ongoing risk knowledge, economic value and competitive differentiation.



(and vice versa). We've found that best practice is to reduce false passes. Unfortunately, false passes are like finding a needle in a haystack. The approach to risk-based judgmental testing is important to find issues in an economical manner.

- **Human Operator Approach:** The machine exception, false positive, and false negative approach were previously outlined. In addition, the human operator/risk reviewer should be actively involved in updating the learning. This could help train the machine on a new document or rule. The accumulated platform learning visual on page 71 demonstrates the iterative nature of the learning process. Great care should be taken to implement and routinize the learning process.[15]

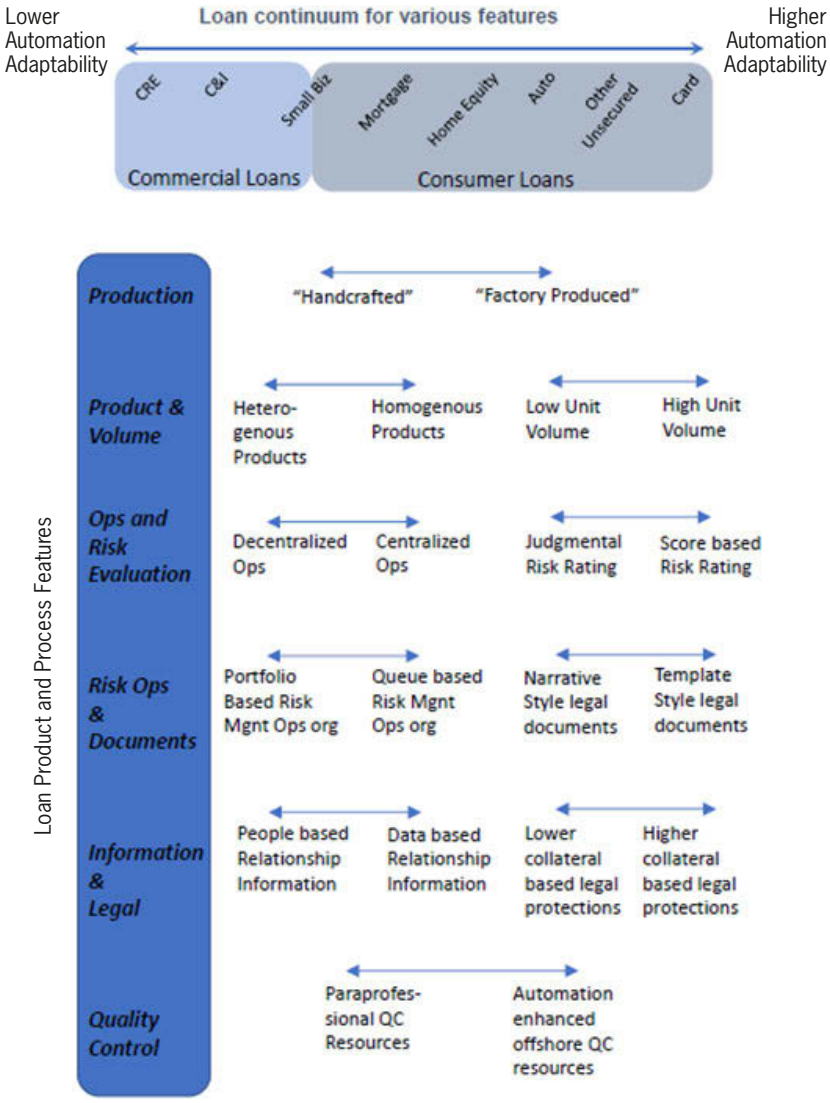
Conclusion

As we've outlined, the use of statistics occurs early in the transactional testing process, when it is determined that a sampling approach is appropriate to draw an inference upon the population. Other types of sampling, like judgmental sampling, are used when an inference for the population is not required. Increasingly, we're seeing the actual testing performed using automated tools. Depending on the kind of automation, statistics may be applied as part of this automation. We've found that there are several important implications to consider based on the use of statistics and automation in the transactional testing process, including:

- A reminder of typical statistics used for making a risk inference.
- How to interpret statistical testing results.
- How the use of statistics in automation is different than traditional sample-based risk testing.
- How to interpret false positives and false negatives for automated testing.
- How to integrate automated and manual testing.
- How to use a framework for machine learning.
- Key adaptability features for different loan types.

Appendix 1 - Automation Adaptability Framework:

The following diagram describes typical loan products most adaptable to automation (on the right side of the axis), as opposed to those least adaptable to automation (to the left). Generally, higher volume, homogenous products will be more adaptable to automation. Below are the loan products and their related features. These features help dimension the products and their relationship to automation adaptability. ®



1. Background of internal controls related to both detective and preventative focus areas: https://www.fa.ufl.edu/directives/types-of-internal-controls/
2. Regulatory guidance on acceptable statistical sampling techniques: https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/sampling-methodologies/pub-ch-sampling-methodologies.pdf
3. Overview of statistical and judgmental sampling approaches and the role of expert judgement in statistical testing: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6474725/
4. An example of related automation technology is IBM's Cloud Pak for Business Automation https://www.ibm.com/cloud/cloud-pak-for-business-automation
5. It is important to determine if the auto-

mated testing has a statistical element. If the data involved in a test is already structured and has been tested for accuracy, it may not have a statistical element. In this case, traditional manual testing protocols maybe used. See Section 3 for manual test outcomes.

6. Background on the latest techniques for ingesting large volumes of PDFs: https://www.ibm.com/blogs/research/2020/05/largest-dataset-for-document-layout-analysis-used-to-ingest-covid-19-data/
7. Background on confidence interval estimation: https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
8. Background and techniques for refining a population to increase accuracy and inference quality: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105563/
9. Sample size example drawn from the OCC: https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/sampling-methodologies/pub-ch-sampling-methodologies.pdf
10. Mathematically, this is because the binomial model assumes large population sizes. Sample size $n = \ln(1 - c) / \ln(1 - p)$, where c is the confidence level and p is the tolerance rate.
11. The article from footnote 8 provides context for tolerance, confidence level, and sample testing interpretation.
12. Lending-focused background regarding the statistical categories of False Negatives and False Positives: https://towardsdatascience.com/model-performance-cost-functions-for-classification-models-a7b1b00ba60
13. A bank that is performing an activity that is not in compliance with one or many regulatory obligations is taking the risk of a regulatory action. A False Negative, by definition, is a risk test that reports to be a pass but is actually a fail. This error of omission could be subject to regulatory action.
14. This is an example case study from a major U.S. bank client as observed in the regulatory risk context.
15. The category description is based on the authors' past client and industry experience.

JEFF HULETT is Director at Promontory Financial Group. He can be reached at jhulett@promontory.com.

RONALD CATHCART is Managing Director at Promontory Financial Group. He can be reached at rcathcart@promontory.com.

CONWAY DODGE is Managing Director at Promontory Financial Group. He can be reached at cdodge@promontory.com.

KRIS MCINTIRE is Managing Director at Promontory Financial Group. He can be reached at kmcintire@promontory.com.